

Hands-on Scikit-Learn for Machine Learning Applications

**Data Science Fundamentals
with Python**

David Paper

Apress®

Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python

David Paper
Logan, UT, USA

ISBN-13 (pbk): 978-1-4842-5372-4
<https://doi.org/10.1007/978-1-4842-5373-1>

ISBN-13 (electronic): 978-1-4842-5373-1

Copyright © 2020 by David Paper

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Jonathan Gennick
Development Editor: Laura Berendson
Coordinating Editor: Jill Balzano

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit <http://www.apress.com/rights-permissions>.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at <http://www.apress.com/bulk-sales>.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484253724. For more detailed information, please visit <http://www.apress.com/source-code>.

Printed on acid-free paper

For my mother, brothers, and friends.

Table of Contents

About the Author ix

About the Technical Reviewer xi

Introduction xiii

Chapter 1: Introduction to Scikit-Learn 1

Machine Learning 1

Anaconda 2

Scikit-Learn 3

Data Sets..... 3

Characterize Data 4

 Simple Classification Data..... 4

 Complex Classification Data 14

 Regression Data 21

Feature Scaling 27

Dimensionality Reduction 30

Chapter 2: Classification from Simple Training Sets 37

Simple Data Sets..... 38

 Classifying Wine Data 38

 Classifying Digits 43

 Classifying Bank Data..... 52

 Classifying make_moons 64

Chapter 3: Classification from Complex Training Sets 71

Complex Data Sets..... 71

 Classifying fetch_20newsgroups 71

 Classifying MNIST..... 79

 Classifying fetch_lfw_people 95

TABLE OF CONTENTS

Chapter 4: Predictive Modeling Through Regression 105

Regression Data Sets..... 105

Regressing tips 106

Regressing boston 117

Regressing wine data 122

Chapter 5: Scikit-Learn Classifier Tuning from Simple Training Sets..... 137

Tuning Data Sets 139

Tuning Iris Data 140

Tuning Digits Data 144

Tuning Bank Data..... 149

Tuning Wine Data 157

Chapter 6: Scikit-Learn Classifier Tuning from Complex Training Sets 165

Tuning Data Sets 166

Tuning fetch_1fw_people 166

Tuning MNIST 175

Tuning fetch_20newsgroups..... 184

Chapter 7: Scikit-Learn Regression Tuning 189

Tuning Data Sets 190

Tuning tips..... 190

Tuning boston..... 199

Tuning wine..... 208

Chapter 8: Putting It All Together 215

The Journey 215

Value and Cost 216

MNIST Value and Cost..... 218

 Explaining MNIST to Money People 222

 Explaining Output to Money People..... 222

 Explaining the Confusion Matrix to Money People 223

Explaining Visualizations to Money People.....	224
Value and Cost.....	224
fetch_lfw_people Value and Cost.....	225
Explaining fetch_lfw_people to Money People.....	229
Explaining Output to Money People.....	229
Explaining Visualizations to Money People.....	230
Value and Cost.....	230
fetch_20newsgroups Value and Cost.....	231
Explaining fetch_20newsgroups to Money People.....	235
Explaining Output to Money People.....	235
Explaining the Confusion Matrix to Money People	235
Value and Cost.....	236
Index.....	239

About the Author



Dr. David Paper is a professor at Utah State University in the Management Information Systems department. He is the author of two books – *Web Programming for Business: PHP Object-Oriented Programming with Oracle* and *Data Science Fundamentals for Python and MongoDB*. He has over 70 publications in refereed journals such as *Organizational Research Methods*, *Communications of the ACM*, *Information & Management*, *Information Resource Management Journal*, *Communications of the AIS*, *Journal of Information Technology Case and Application Research*, and *Long Range Planning*. He has also served on several editorial boards in various capacities, including associate editor. Besides

growing up in family businesses, Dr. Paper has worked for Texas Instruments, DLS, Inc., and the Phoenix Small Business Administration. He has performed IS consulting work for IBM, AT&T, Octel, Utah Department of Transportation, and the Space Dynamics Laboratory. Dr. Paper's teaching and research interests include data science, machine learning, process reengineering, object-oriented programming, and change management.

About the Technical Reviewer



Jojo Moolayil is an artificial intelligence, deep learning, machine learning, and decision science professional and published author of three books: *Smarter Decisions – The Intersection of Internet of Things and Decision Science*, *Learn Keras for Deep Neural Networks*, and *Applied Supervised Learning with R*. He has worked with industry leaders on several high-impact and critical data science and machine learning projects across multiple verticals. He is currently associated with Amazon Web Services as a research scientist – AI.

Jojo was born and raised in Pune, India, and graduated from the University of Pune with a major in Information Technology Engineering. He started his career with Mu Sigma Inc., the world's largest pure-play analytics provider, and worked with the leaders of many Fortune 50 clients. He later worked with Flutura – an IoT analytics start-up – and GE, the pioneer and leader in Industrial AI.

He currently resides in Vancouver, BC. Apart from authoring books on deep learning, decision science, and IoT, Jojo has also been a technical reviewer for various books on the same subject with Apress and Packt publications. He is an active Data Science tutor and maintains a blog at <http://blog.jojomoolayil.com>.

- Jojo's personal web site: www.jojomoolayil.com
- Business e-mail: mail@jojomoolayil.com

Introduction

We apply the popular Scikit-Learn library to demonstrate machine learning exercises with Python code to help readers solve machine learning problems. The book is designed for those with intermediate programming skills and some experience with machine learning algorithms. We focus on application of the algorithms rather than theory. So, readers should read about the theory online or from other sources if appropriate. The reader should also be willing to spend a lot of time working through the code examples because they are pretty deep. But, the effort will pay off because the examples are intended to help the reader tackle complex problems.

The book is organized into eight chapters. Chapter 1 introduces the topic of machine learning, Anaconda, and Scikit-Learn. Chapters 2 and 3 introduce algorithmic classification. Chapter 2 classifies simple data sets and Chapter 3 classifies complex ones. Chapter 4 introduces predictive modeling with regression. Chapters 5 and 6 introduce classification tuning. Chapter 5 tunes simple data sets and Chapter 6 tunes complex ones. Chapter 7 introduces predictive modeling regression tuning. Chapter 8 puts all knowledge together to review and present findings in a holistic manner.

Download this book's example data by clicking the Download source code button found on the book's catalog page at <https://www.apress.com/us/book/9781484253724>.